Interrelation between validity indices and validation in French administrative data

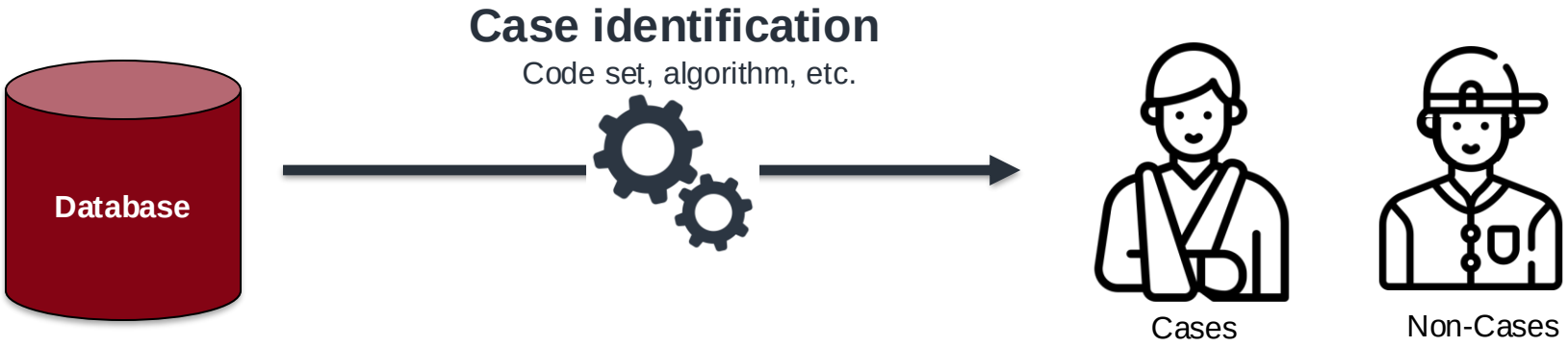
Validation and Validation of Validation

Workshop, ARS Toscana, Florence Italy, 27 March 2025

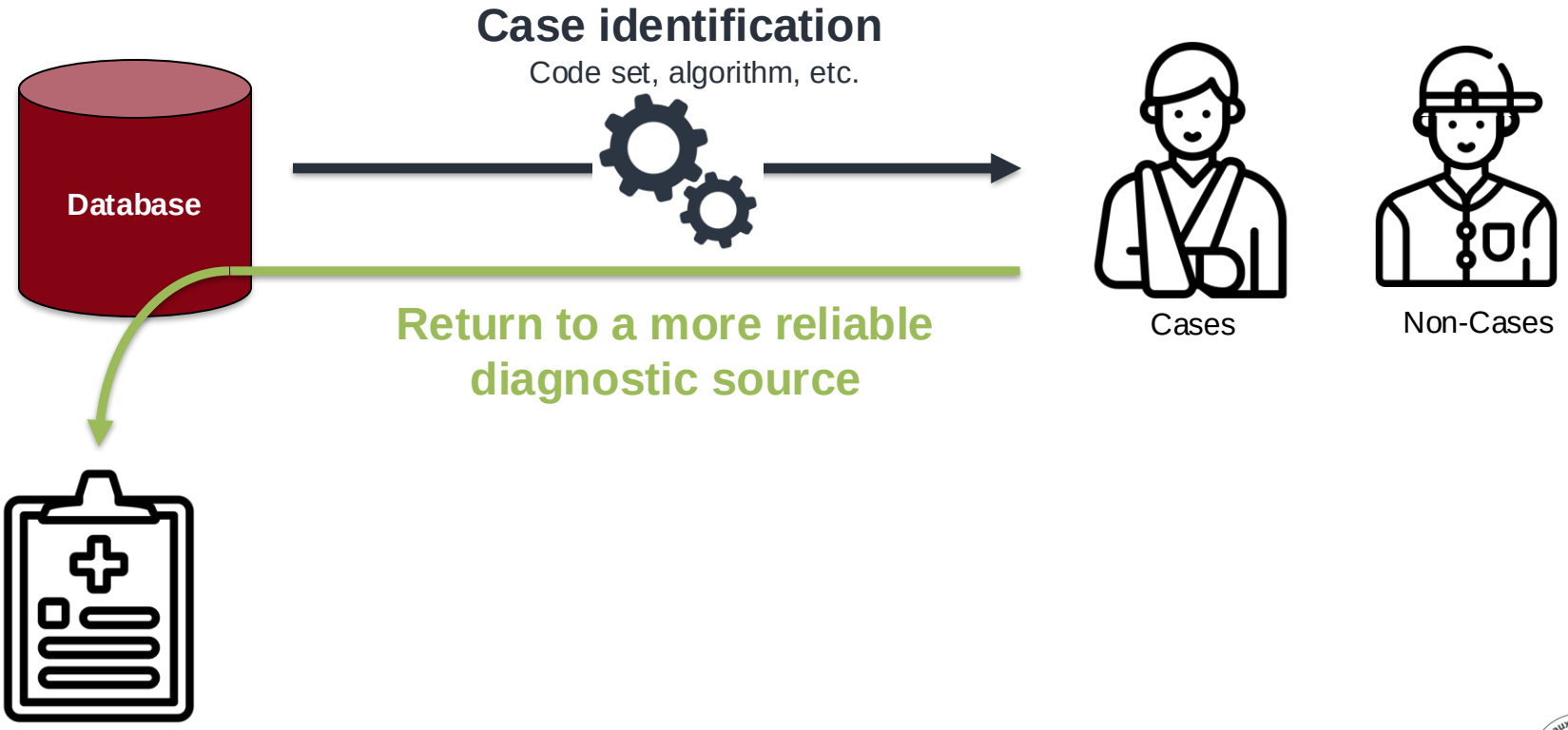
Nicolas THURIN

How to conduct a validation study

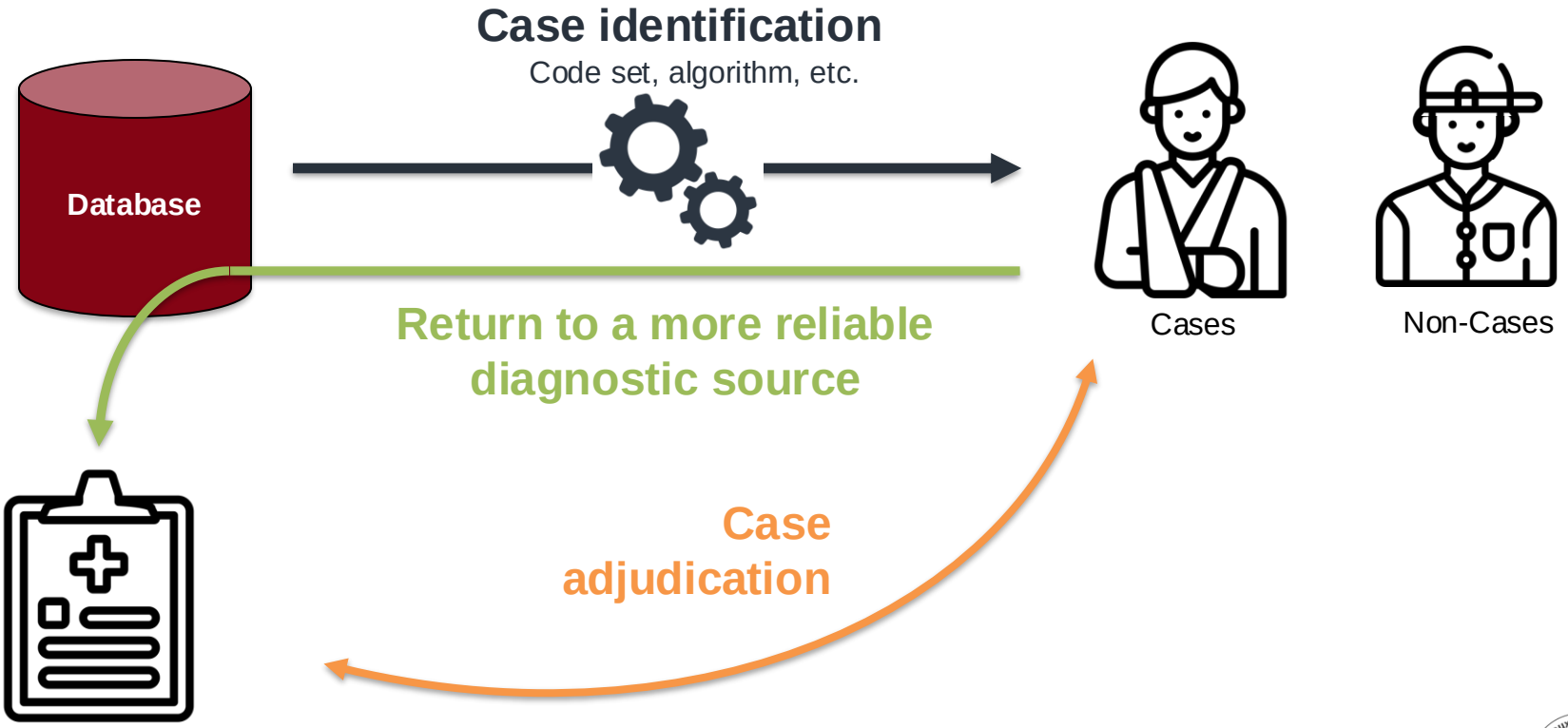
How to conduct a validation study



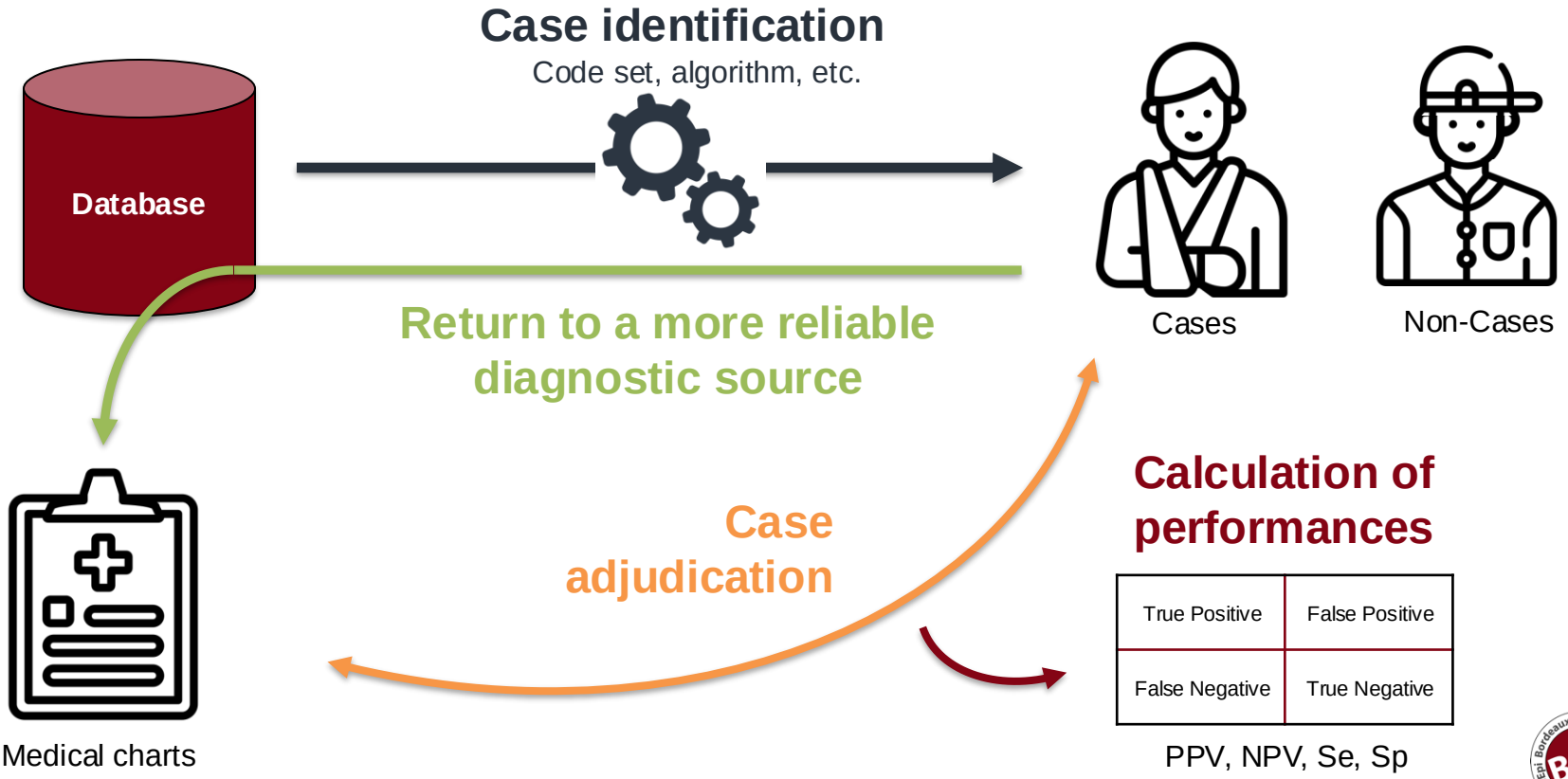
How to conduct a validation study



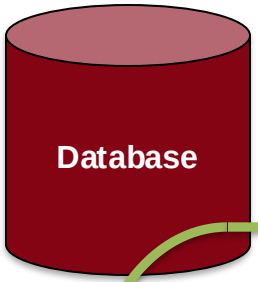
How to conduct a validation study

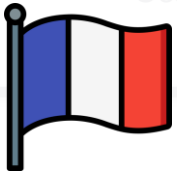


How to conduct a validation study



How to conduct a validation study





Patient consent is required

Cases

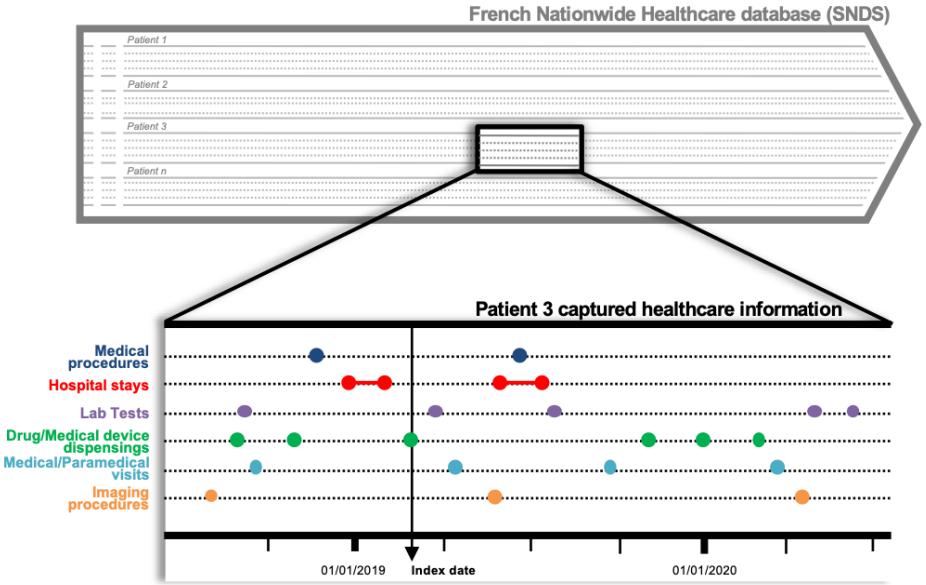
Non-Cases

Return to a more reliable diagnostic source

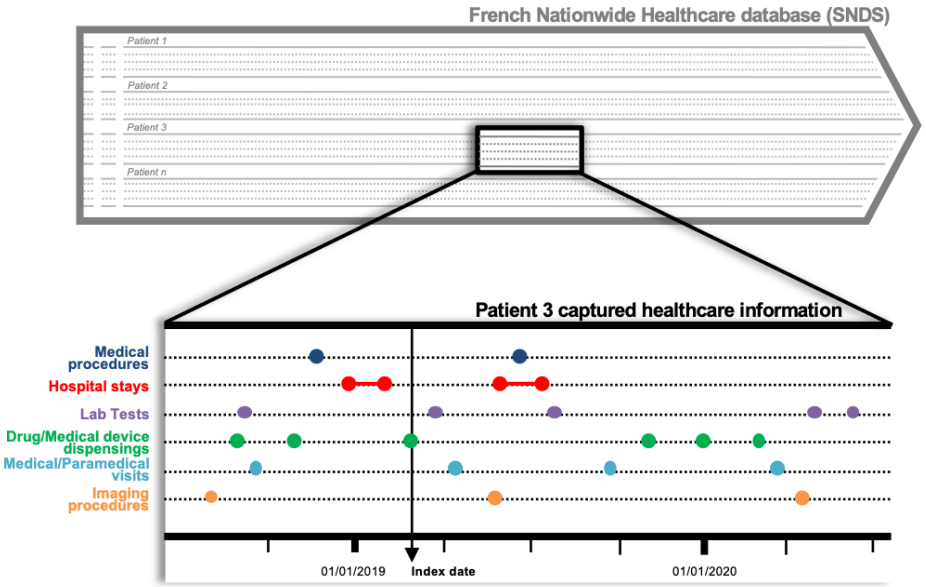
Medical charts
Gold standard, etc.

What alternative do we have?

The power of claims



The power of claims

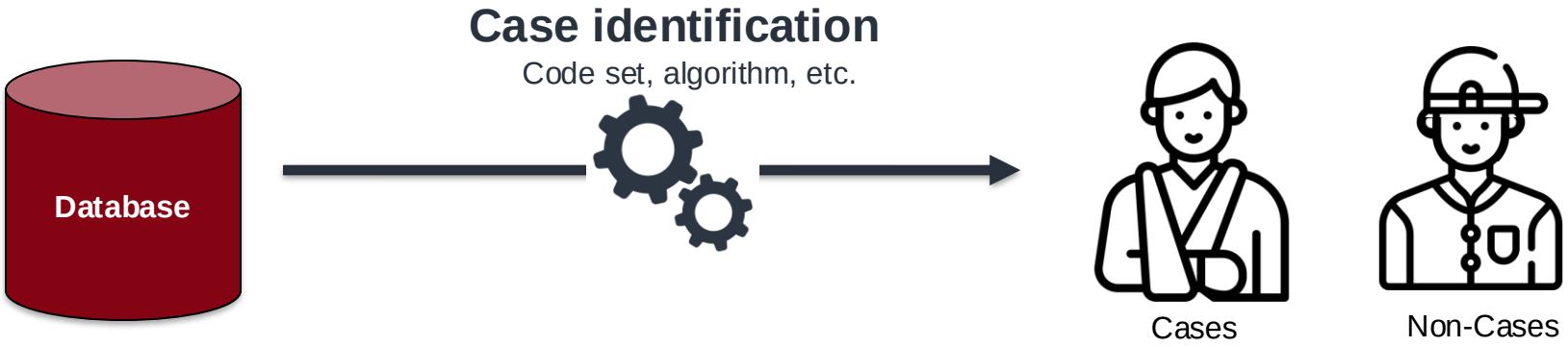


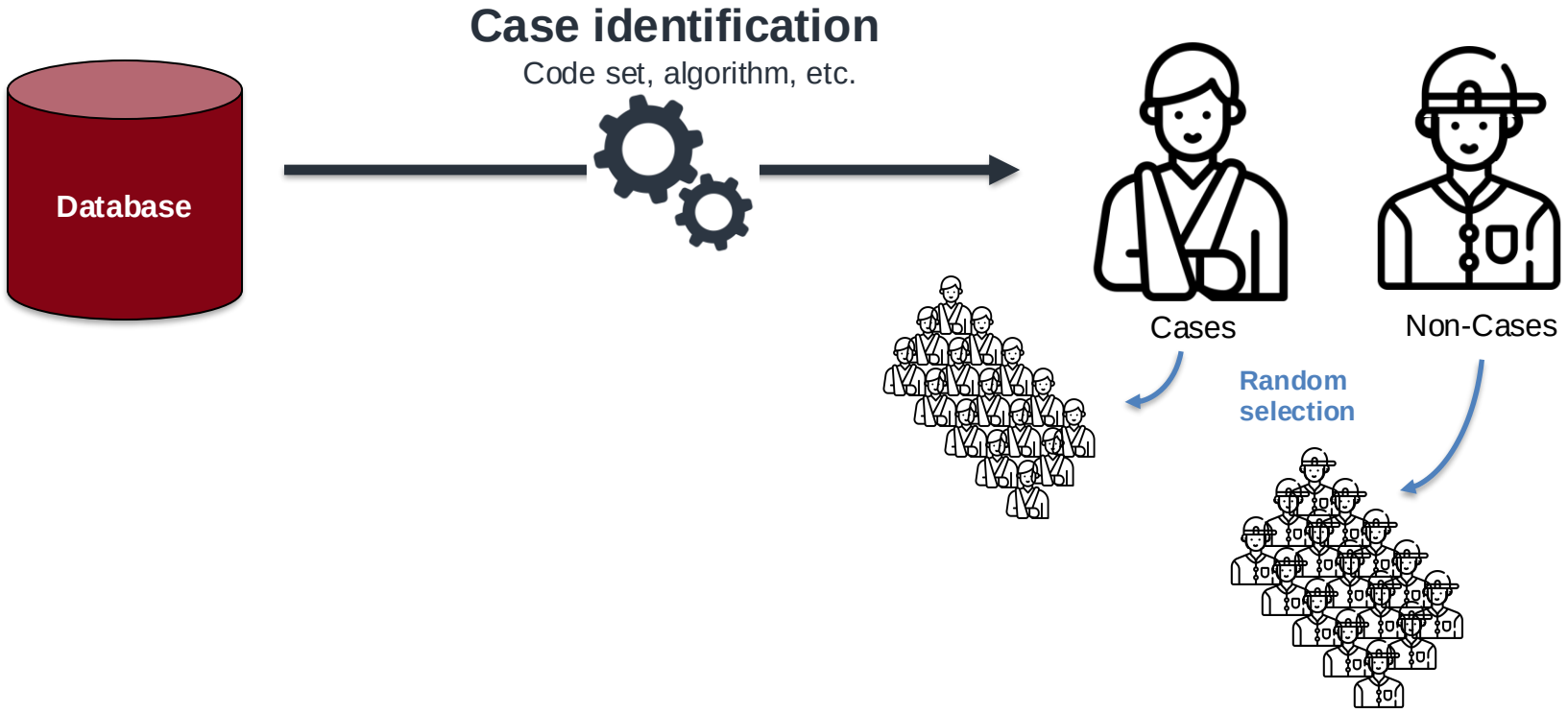
Anonymized Reconstituted Electronic Health Records (rEHR) of Patient n°3

Year of inclusion (index year): 05/02/2019 **Age class at the inclusion:** [65 - 70[years **Gender:** Male

Type of healthcare encounter	Encounter start date	Encounter end date	Description
...	...	...	...
Drug dispensing	-11 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Bisoprolol 5 MG CPR 30 - ATC C07BB07*
Lab test	-5 days		Code and description of the lab test *e.g. Blood count including platelet*
Hospital stay	0	+ 5 days	ICD-10 principal (PD), related (RD) and associated (AD) discharge diagnoses *e.g. PD I26. Pulmonary embolism*
Imaging procedure	+1 day		Code and description of the imaging procedure *e.g. EJQM003 - Doppler ultrasound of lower limb and iliac veins for deep vein thrombosis*
Drug dispensing	+6 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Eliquis 5 MG CPR 60 - ATC B01AF02*
Medical visit	+20 ays		Physician specialty *e.g. General practitioner*
Drug dispensing	+30 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Bisoprolol 5 MG CPR 30 - ATC C07BB07*
Drug dispensing	+30 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Eliquis 5 MG CPR 60 - ATC B01AF02*
Medical procedure	+1 months 2 days		Code and description of the medical procedure *e.g. NZQJ001 - Unilateral or bilateral remnography of lower limb segments, with injection of contrast agent*
...	...	...	...

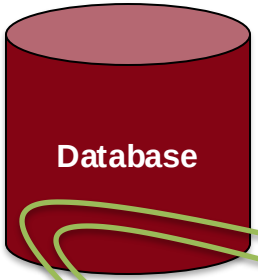
Data
de-identification


Case identification
Code set, algorithm, etc.
Database
Cases
Non-Cases


Case identification
Code set, algorithm, etc.
Database
Cases
Non-Cases
Random selection

Case identification

Code set, algorithm, etc.

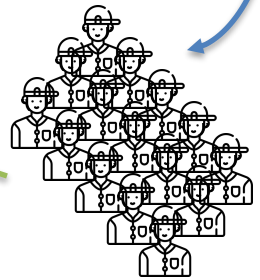


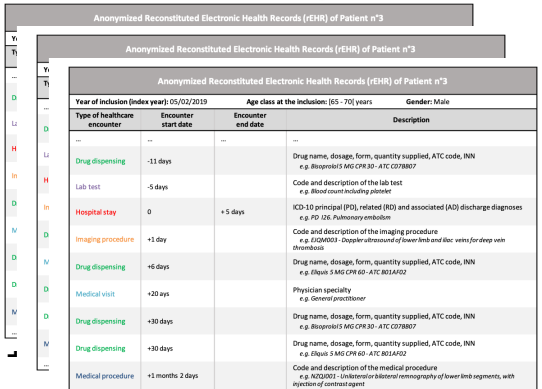
Cases

Non-Cases

Reconstitution of anonymised EHR

Random selection





Anonymized Reconstituted Electronic Health Records (rEHR) of Patient n°3

Anonymized Reconstituted Electronic Health Records (rEHR) of Patient n°3

Anonymized Reconstituted Electronic Health Records (rEHR) of Patient n°3

Year of inclusion (index year): 05/02/2019 **Age class at the inclusion:** [65 - 70[years **Gender:** Male

Type of healthcare encounter	Encounter start date	Encounter end date	Description
...	...	...	...
Drug dispensing	-11 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Bisoprolol 5 MG CPR 30 - ATC C07AB07*
Lab test	-5 days		Code and description of the lab test *e.g. Blood count including platelet*
Hospital stay	0	+ 5 days	ICD-10 principal (PD), related (RD) and associated (AD) discharge diagnoses *e.g. PD I26. Pulmonary embolism*
Imaging procedure	+1 day		Code and description of the imaging procedure *e.g. EJQM003 - Doppler ultrasound of lower limb and iliac veins for deep vein thrombosis*
Drug dispensing	+6 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Eliquis 5 MG CPR 60 - ATC B01AF02*
Medical visit	+20 ays		Physician specialty *e.g. General practitioner*
Drug dispensing	+30 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Bisoprolol 5 MG CPR 30 - ATC C07AB07*
Drug dispensing	+30 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Eliquis 5 MG CPR 60 - ATC B01AF02*
Medical procedure	+1 months 2 days		Code and description of the medical procedure *e.g. NZQJ001 - Unilateral or bilateral remnography of lower limb segments, with injection of contrast agent*

Case identification

Code set, algorithm, etc.

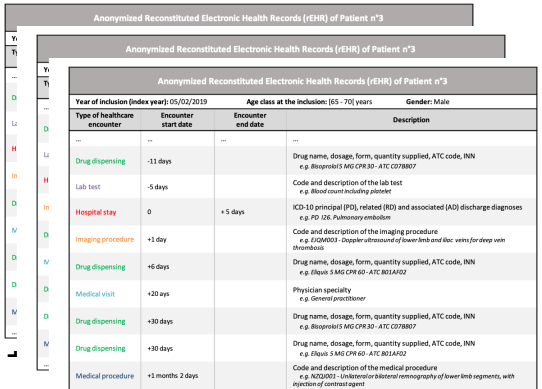
Database

Cases

Non-Cases

Reconstitution of anonymised EHR

Random selection



Anonymized Reconstituted Electronic Health Records (rEHR) of Patient n°3

Year of inclusion (index year): 05/02/2019 Age class at the inclusion: [65 - 70[years Gender: Male

Type of healthcare encounter	Encounter start date	Encounter end date	Description
...	...	...	...
Drug dispensing	-11 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Bisoprolol 5 MG CPR 30 - ATC C07AB07*
Lab test	-5 days		Code and description of the lab test *e.g. Blood count including platelet*
Hospital stay	0	+ 5 days	ICD-10 principal (PD), related (RD) and associated (AD) discharge diagnoses *e.g. PD: I26. Pulmonary embolism*
Imaging procedure	+1 day		Code and description of the imaging procedure *e.g. EIQM003 - Doppler ultrasound of lower limb and iliac veins for deep vein thrombosis*
Drug dispensing	+6 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Eliquis 5 MG CPR 60 - ATC B01AF02*
Medical visit	+20 ays		Physician specialty *e.g. General practitioner*
Drug dispensing	+30 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Bisoprolol 5 MG CPR 30 - ATC C07AB07*
Drug dispensing	+30 days		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. Eliquis 5 MG CPR 60 - ATC B01AF02*
Medical procedure	+1 months 2 days		Code and description of the medical procedure *e.g. NZQJ001 - Unilateral or bilateral remnography of lower limb segments, with injection of contrast agent*

Blinded case adjudication by physicians

True Positive	False Positive
False Negative	True Negative

Validity indices

Table 1. Validity indices for dichotomous data: Sensitivity (SE), specificity (SP) positive (PPV) and negative predictive value (NPV) the observed (P) and true prevalence (π).

		'Gold' standard		
		Positive	Negative	Validity index
Case Finding Algorithm	Positive	Nr. of True positives TP	Nr. of False positives FP	PPV = TP/(TP+FP)
	Negative	Nr. of False negatives FN	Nr. of True negatives TN	NPV = TN/(FN+TN)
	Validity index	SE = TP/(TP + FN)	SP = TN/(FP + TN)	N = TP+FP+FN+TN P = (TP+FP)/N π = (TP+FN)/N

Table 2. Overview of the interrelations between validity indices and the true prevalence, given the observed prevalenceP and two other parameters.

10.	P, PPV, NPV	$\Pi = (1-\mathrm{P})(1-\mathrm{NPV}) + \mathrm{P} \times \mathrm{PPV}$	$\mathrm{SE} = \frac{\mathrm{P}\times\mathrm{PPV}}{(1-\mathrm{P})(1-\mathrm{NPV})+\mathrm{P}\times\mathrm{PPV}}$	$\mathrm{SP} = \frac{(1-\mathrm{P})\times\mathrm{NPV}}{1-(\mathrm{P}\times\mathrm{PPV}+(1-\mathrm{P})(1-\mathrm{NPV}))}$

From: Bollaerts, K., Rekkas, A., De Smedt, T., Dodd, C., Andrews, N., & Gini, R. (2020). Disease misclassification in electronic healthcare database studies: Deriving validity indices-A contribution from the ADVANCE project. *PloS one*, *15*(4), e0231333. https://doi.org/10.1371/journal.pone.0231333

Conditions

- The outcome of interest must be managed by a specific sequence of care and encounters

- The considered healthcare database must capture in an exhaustive way a sufficient number of medical elements in line with the outcome of interest

Applications of intra-database validation

- Metastatic castration-resistant prostate cancer in the SNDS[1]
- Relapse in multiple sclerosis in the SNDS[1]
- Venous thromboembolism in the SNDS[2]
- Acute Myocardial Infarction, Stroke, and Cardiovascular Death in German Health Insurance Data[3]

1. Thurin, N. H., Bosco-Levy, P., Blin, P., et al. (2021). Intra-database validation of case-identifying algorithms using reconstituted electronic health records from healthcare claims data. BMC medical research methodology, 21(1), 95. https://doi.org/10.1186/s12874-021-01285-y
2. Thurin, N. H., Grelaud, A., Grolleau, A., et al. (2024). Design and validation of algorithms to identify venous thromboembolism in the French National Healthcare Database. Pharmacoepidemiology and drug safety, 33(4), e5781. https://doi.org/10.1002/pds.5781
3. Platzbecker, K., Voss, A., Reinold, J., et al. (2022). Validation of Algorithms to Identify Acute Myocardial Infarction, Stroke, and Cardiovascular Death in German Health Insurance Data. Clinical epidemiology, 14, 1351–1361. https://doi.org/10.2147/CLEP.S380314

Illustration

- Metastatic castration-resistant prostate cancer (SNDS)
  - PPV = 97% (95%CI = [93; 100])
  - NPV = 99% (95%CI = [97; 100])
- Relapse in multiple sclerosis (SNDS)
  - PPV = 95% (95%CI = [91; 99])
  - NPV=100%
- Venous thromboembolism (SNDS)
  - PPV = 92% (95%CI = [86; 98])
- Acute Myocardial Infarction (GePaRD)
  - PPV = 97.6% (95%CI = [94.8; 99.1])
- Stroke (GePaRD)
  - PPV = 94.8% (95%CI = [91.3; 97.2])

Validation of Validation

Illustration

- Metastatic castration-resistant prostate cancer (SNDS)
  - PPV = 97% (95%CI = [93; 100])
  - NPV = 99% (95%CI = [97; 100])
- **Relapse in multiple sclerosis (SNDS)**
  - **PPV = 95% (95%CI = [91; 99])**
  - **NPV=100%**
- Venous thromboembolism (SNDS)
  - PPV = 92% (95%CI = [86; 98])
- Acute Myocardial Infarction (GePaRD)
  - PPV = 97.6% (95%CI = [94.8; 99.1])
- Stroke (GePaRD)
  - PPV = 94.8% (95%CI = [91.3; 97.2])

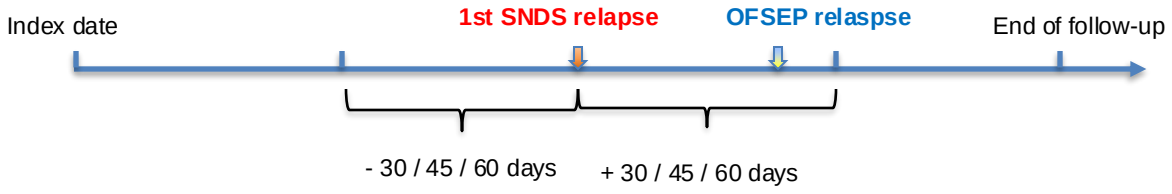
Relapse in multiple sclerosis (MS)

- Algorithm designed in a study to assess the comparative effectiveness of Dimethyl fumarate on the onset of MS relapse
  - Data source = SNDS
  - Population = MS patients
  - Algorithm based on
    - hospital stays ± high dose corticosteroids
    - Dispensing of high-dose corticosteroids in outpatient settings

- PPV = 95% (95%CI = [91; 99]), NPV=100%

Validation of the validation

- Prospective clinical database of the French Multiple Sclerosis Observatory (OFSEP)
  - MS-dedicated EHRs populated by a neurologist following a patient visit
  - Information can be retrospective
  - Clinical evaluation
  - Neurological episodes
- Linkage with the SNDS
- Replication of the population used for the dimethyl fumarate study
- Assessment of the consistency between the algorithm-identified cases and OFSEP-recorded cases (±30 days, ±45 days, ±60 days)



Results

	Initial algorithm *vs. OFSEP*			Initial algorithm *vs. SNDS*
	+/- 30j	+/- 45j	+/- 60j	
PPV	60.9	**65.5**	**68.6**	**95**
NPV	89.8	**89.8**	**89.8**	**100**
Se	72.1	74.1	75.0	
Spe	83.8	85.3	86.5	
Precision	80.7	82.3	83.2	

Discussion

Discussion

- Although selected on the same inclusion and exclusion criteria, the assessed population may differ

Discussion

- Although selected on the same inclusion and exclusion criteria, the assessed population may differ
- **Who is right?**

Discussion

- Although selected on the same inclusion and exclusion criteria, the assessed population may differ
- **Who is right?**
  - SNDS: prospective database collecting the exhaustivity of reimbursed healthcare encounters
  - OFSEP: French Multiple Sclerosis Observatory with detailed clinical information

Discussion

- Although selected on the same inclusion and exclusion criteria, the assessed population may differ
- **Who is right?**
  - SNDS: prospective database collecting the exhaustivity of reimbursed healthcare encounters
  - OFSEP: French Multiple Sclerosis Observatory with detailed clinical information
- Is a high-dose corticosteroid dispensing in the SNDS always associated with a relapse (dispensing “in case of”)?
- Do all patients with a relapse attend a neurologist and report their episodes?

Discussion

- What is the impact of such uncertainty on potential study results?
- Depend on
  - What is this algorithm used for (inclusion criteria *vs.* outcome identification)
  - Study type/objective (burden of disease *vs.* comparative analysis)
  - Potential expected bias (what are we missing in our data?)

Conclusion

Conclusion

- Standard methodology may not always be applicable
  - Please be creative and find a way to validate your work

Conclusion

- Standard methodology may not always be applicable
  - Please be creative and find a way to validate your work

- Is your gold standard really gold?
  - Please carefully assess the gold standard that you will be considering for your validation study
  - Please carefully consider the impact of potential bias on your results